

뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

# Welcome To News Bot !

2020.12.11 (Fri)

참여자 - 4

김종찬

서기현

유승균

이기중

참여자 - 4

- 김종찬
- 서기현
- 유승균
- 이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

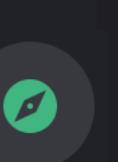
## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트



#NewsBot에 메시지 보내기





NEWS NEWS  
SAMPLE NEWS NEWS  
NEWS NEWS  
NEWS NEWS  
뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

-  김종찬
-  서기현
-  유승균
-  이기중

+ #NewsBot에 메시지 보내기





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

서기현

유승균

이기중

## # 프로젝트 동기

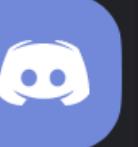
“내 관심분야에 관한 뉴스 기사만 모아서 동향 파악을 하고 싶은데 매일 검색하기 좀 귀찮음...”

“검색을 하면 내가 보고자 했던 뉴스 내용이 아닌 것들도 많아..”

“내가 보고 싶은 뉴스들만 모아서 보기 좋게 메신저로 보내줬으면 좋겠다!”

## # 프로젝트 목표

“뉴스 챗봇을 만들어 나의 귀차니즘도 해결하고, 매일 관심분야의 새로운 소식들을 얻어보자!”



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

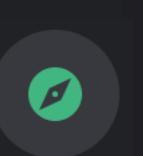
- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

- 김종찬
- 서기현
- 유승균
- 이기중

+ #NewsBot에 메시지 보내기





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

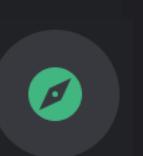
서기현

유승균

이기중

+ #NewsBot에 메시지 보내기





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

# How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

서기현

유승균

이기중

## # 프로젝트 동기

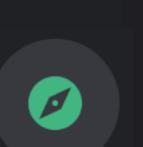
“내 관심분야에 관한 뉴스 기사만 모아서 동향 파악을 하고 싶은데 매일 검색하기 좀 귀찮음...”

“검색을 하면 내가 보고자 했던 뉴스 내용이 아닌 것들도 많아..”

“내가 보고 싶은 뉴스들만 모아서 보기 좋게 메신저로 보내줬으면 좋겠다!”

## # 프로젝트 목표

“뉴스 챗봇을 만들어 나의 귀차니즘도 해결하고, 매일 관심분야의 새로운 소식들을 얻어보자!”



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

서기현

유승균

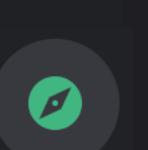
이기중

+ #NewsBot에 메시지 보내기



GIF





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

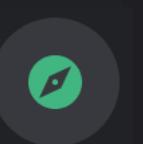
-  김종찬
-  서기현
-  유승균
-  이기중

+ #NewsBot에 메시지 보내기



GIF





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

서기현

유승균

이기중

## # 프로젝트 동기

“내 관심분야에 관한 뉴스 기사만 모아서 동향 파악을 하고 싶은데 매일 검색하기 좀 귀찮음...”

“검색을 하면 내가 보고자 했던 뉴스 내용이 아닌 것들도 많아..”

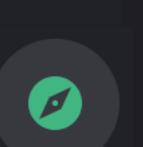
“내가 보고 싶은 뉴스들만 모아서 보기 좋게 메신저로 보내줬으면 좋겠다!”

## # 프로젝트 목표

“뉴스 챗봇을 만들어 나의 귀차니즘도 해결하고, 매일 관심분야의 새로운 소식들을 얻어보자!”

+ #NewsBot에 메시지 보내기





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

서기현

유승균

이기중

## # 프로젝트 동기

“내 관심분야에 관한 뉴스 기사만 모아서 동향 파악을 하고 싶은데 매일 검색하기 좀 귀찮음...”

“검색을 하면 내가 보고자 했던 뉴스 내용이 아닌 것들도 많아..”

“내가 보고 싶은 뉴스들만 모아서 보기 좋게 메신저로 보내줬으면 좋겠다!”

## # 프로젝트 목표

“뉴스 챗봇을 만들어 나의 귀차니즘도 해결하고, 매일 관심분야의 새로운 소식들을 얻어보자!”



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

참여자 - 4

김종찬

서기현

유승균

이기중

+ #NewsBot에 메시지 보내기





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

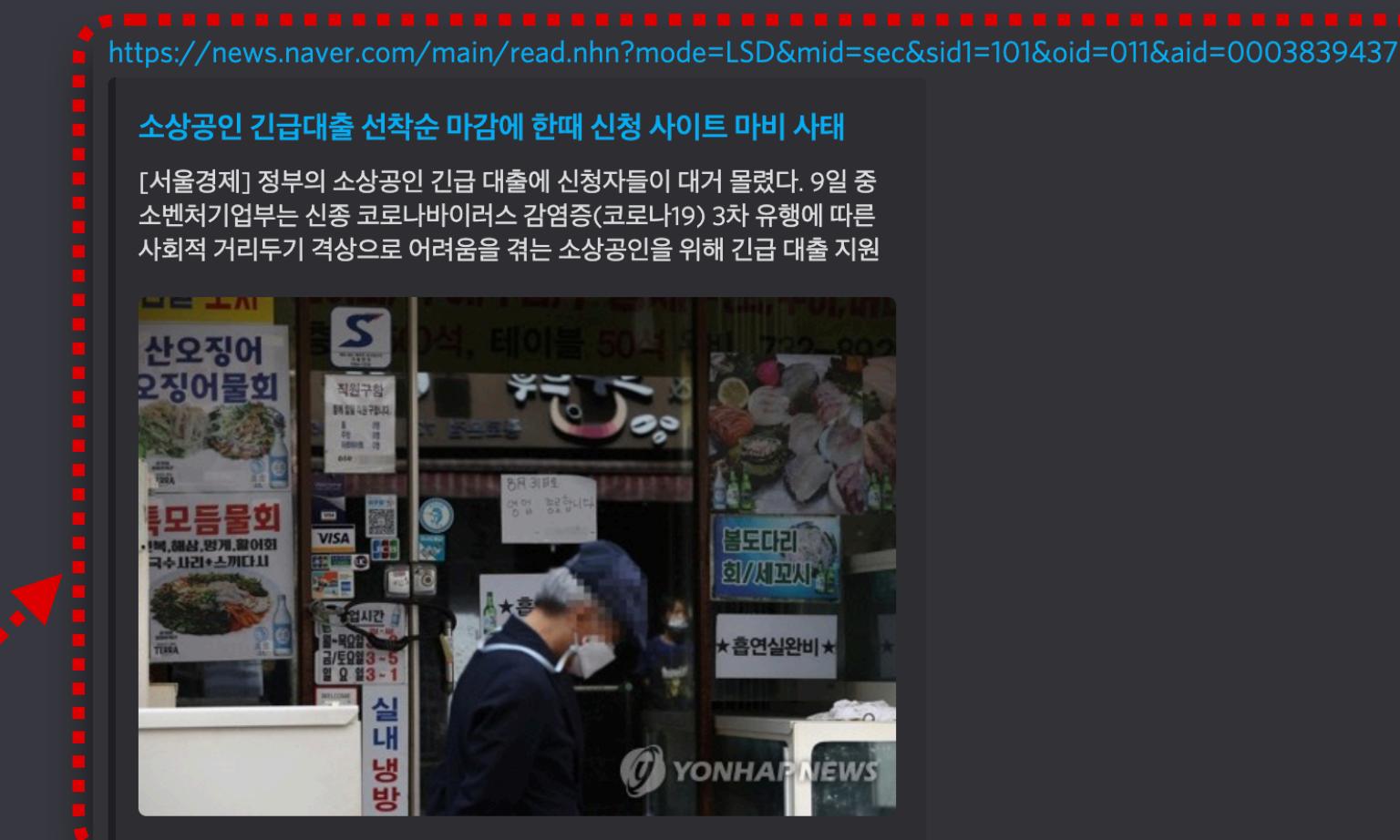
# 결과

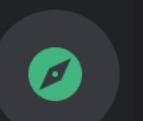
## # How?

- ① 데이터 수집 (Scrapy 크롤링)
- ② 데이터 저장 및 관리 (csv, MongoDB)
- ③ 챗봇 개발 (NLP / WordCloud)
- ④ 챗봇 테스트

## 결과 예시

+ !content 2020.12.07 인공지능 코로나





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 url 수집 (2019.01.01 ~ 2020.11.29)

```
# 프로젝트 시작
!scrapy startproject get_url

# Items.py
%%writefile get_url/get_url/items.py
import scrapy

class GetUrlItem(scrapy.Item):
    date = scrapy.Field()
    categ = scrapy.Field()
    last_p = scrapy.Field()

# Spider.py
%% writefile
get_url / get_url / spiders / spider.py
import scrapy

from .naver_articles import *
from scrapy.http import TextResponse
from get_url.items import GetUrlItem
```

24 4 4

참여자—4

김종찬

서기현

유승균

이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 url 수집 (2019.01.01 ~ 2020.11.29)

```
class GetUrlSpider(scrapy.Spider):
    name = 'get_url'
    allow_domain = ["https://news.naver.com"]
    user_agent = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.198 Safari/537.36'

    def start_requests(self):
        dates = get_day_list('2019/01/01', '2020/11/29')
        categ_s = [101, 102, 103, 105]
        # dates = list(divide_list(dates, 60))
        for date in dates:
            for categ in categ_s:
                for page in range(1, 250, 10):
                    # 마지막 페이지로
                    url = 'https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&listType=title&sid1={}&date={}&page={}'.format(
                        categ, date, page)
                    yield scrapy.Request(url, callback=self.parse)

    def parse(self, resp):
        item = GetUrlItem()
        try:
            chk_next = resp.xpath('//div[@class="paging"]/a[@class="next nclicks(fls.page)"]/text()')[0].extract()
        except:
            chk_next = '끝'

        if chk_next == '끝':
            pages = resp.xpath('//a[@class="nclicks(fls.page)"]/text() | \
                //*[@id="main_content"]/div[@class="paging"]/strong/text()').extract()
            current_page = resp.url.split('page=')[1]
            if int(current_page) < int(pages[-1]):
                item['date'] = resp.url.split('date=')[1].split('&')[0]
                item['categ'] = resp.url.split('sid1=')[1].split('&')[0]
                item['last_p'] = pages[-1]

        yield item
```

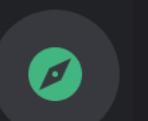
참여자—4

김종찬

서기현

유승균

이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 (2019.01.01 ~ 2020.11.29)

```
!scrapy startproject naver_news_02
%%writefile naver_news_02/naver_news_02/items.py
import scrapy

class NaverNews02Item(scrapy.Item):
    date = scrapy.Field()
    press_agency = scrapy.Field()
    category = scrapy.Field()
    link = scrapy.Field()
    title = scrapy.Field()
    content = scrapy.Field()

%% writefile naver_news_02 / naver_news_02 / spiders / spider.py
# %Load naver_news_02/naver_news_02/spiders/spider.py
import pandas as pd
import time
import re
import requests
import scrapy
from scrapy.http import TextResponse
from naver_news_02.items import NaverNews02Item
```

46 10 ▲

참여자—4

김종찬

서기현

유승균

이기중

## Naver 뉴스 기사 수집 (2019.01.01 ~ 2020.11.29)

```
class NaverNews02Spider(scrapy.Spider):
    name = 'naver_news_02'
    allow_domain = ["https://news.naver.com"]
    user_agent = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.198 Safari/537.36'
    cateq = { # '정치': '100',
        '101': '경제',
        '102': '사회',
        '103': '생활/문화',
        # '세계': '104',
        '105': 'IT/과학'}
```

```
def start_requests(self):
    # df = pd.read_csv('article_url_1.csv')
    df = pd.read_csv('naver_news_02/article_soci.csv')
    rows = df.iloc
    date_ex = '202011'

    for row in rows:
        date_ = str(row['date'])
        date_ = str(date_)[0:7]
        if date_ != date_ex:
            time.sleep(2)
        # print(row['cateq'], row['date'], row['last_p'])
        for page in range(1, int(row['last_p']) + 1):
            url = 'https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&listType=title&sid1={}&date={}&page={}'.format(
                row['cateq'], row['date'], page)
            yield scrapy.Request(url, callback=self.parse)
    date_ex = str(date_)[0:7]
```

참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중

뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index

+

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 (2019.01.01 ~ 2020.11.29)

```

def parse(self, resp):
    links = resp.xpath('//*[@id="main_content"]/div[2]/ul/li/a/@href').extract()

    # links = [resp.urljoin(link) for link in links]
    for link in links:
        yield scrapy.Request(link, callback=self.parse_content)

def parse_content(self, resp):
    item = NaverNews02Item()
    title = resp.xpath('//*[@id="articleTitle"]/text() | //*[@id="content"]/div[1]/div/h2/text() | \
        //h4[@class="title"]/text()')[0].extract()
    date = resp.xpath('//*[@id="main_content"]/div[1]/div[3]/div/span[@class="t11"]/text() | \
        //div[@class="article_info"]/span[@class="author"]/em/text() | \
        //div[@class="info"]/span[1]/text()')[0].extract()
    content = resp.xpath('//*[@id="articleBodyContents"]/text() | \
        //*[@id="articleBodyContents"]/strong/text() | \
        //*[@id="articleBodyContents"]/div/text() | \
        //*[@id="articleBodyContents"]/div/div/text() | \
        //*[@id="articleBodyContents"]/font/text() | \
        //*[@id="articleBodyContents"]/div[2]/ul/li/span/span/text() | \
        //*[@id="newsEndContents"]/text() | \
        //*[@id="articeBody"]/text()').extract()
    content = [text.replace('\xa0', ' ').strip() for text in content]
    categ_num = resp.url.split('sid1=')[1].split('&')[0]

    item['date'] = re.findall('[0-9]{4}[.][0-9]{2}[.][0-9]{2}', date)[0]
    item['category'] = self.categ[categ_num]
    item['press_agency'] = resp.xpath('//a[@class="nclicks(atp_press)"]/img/@title | //div[@class="press_logo"]/a/img/@alt | \
        //*[@id="pressLogo"]/a/img/@alt')[0].extract()
    item['link'] = resp.url
    item['title'] = title.strip()
    item['content'] = '\n'.join(content).strip()

    yield item

```

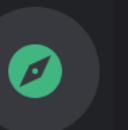
참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중

뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

| Index |              |
|-------|--------------|
| #     | 프로젝트 동기 및 목표 |
| #     | 진행 과정        |
| —     | 데이터 수집       |
| —     | 데이터 저장 및 관리  |
| —     | 챗봇 개발        |
| —     | 챗봇 테스트       |
| #     | 결과           |



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index

+

# 프로젝트 초기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 (2019.01.01 ~ 2020.11.29)

```
%%writefile naver_news_02/naver_news_02/pipelines.py
# %Load naver_news_02/naver_news_02/pipelines.py

from itemadapter import ItemAdapter
from ..mongodb import collection

class NaverNews02Pipeline:
    def process_item(self, item, spider):
        # time.sleep(10)
        data = {
            'p_date': item['date'],
            'category': item['category'],
            'press_agency': item['press_agency'],
            'link': item['link'],
            'title': item['title'],
            'content': item['content'],
        }
        print('*'*5)
        collection.insert(data)
        return item

%%writefile naver_news_02/naver_news_02/mongodb.py
import pymongo
# DB와 연결
client = pymongo.MongoClient('mongodb://127.0.0.1:27017/')
# DB Table 지정
db = client.news
collection = db.articles_society
```

참여자—4

김종찬

서기현

유승균

이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 (2019.01.01 ~ 2020.11.29)

```
%%writefile naver_news_02/naver_news_02/settings.py

BOT_NAME = 'naver_news_02'

SPIDER_MODULES = ['naver_news_02.spiders']
NEWSPIDER_MODULE = 'naver_news_02.spiders'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
#USER_AGENT = 'naver_news_02 (+http://www.yourdomain.com)'

# Obey robots.txt rules
ROBOTSTXT_OBEY = False

# Configure maximum concurrent requests performed by Scrapy (default: 16)
#CONCURRENT_REQUESTS = 32

ITEM_PIPELINES = {
    'naver_news_02.pipelines.NaverNews02Pipeline': 300,
}
```

참여자—4

김종찬

서기현

유승균

이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 자동화 (2020.12.05 ~ )

```
!scrapy startproject naver_news
# Items.py
%%writefile naver_news/naver_news/items.py
# %Load naver_news/naver_news/items.py
import scrapy

class NaverNewsItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    date = scrapy.Field()
    press_agency = scrapy.Field()
    category = scrapy.Field()
    link = scrapy.Field()
    title = scrapy.Field()
    content = scrapy.Field()

    # url 저장하는 함수
%%writefile naver_news/naver_news/spiders/naver_articles.py
import requests
import scrapy
from scrapy.http import TextResponse
from datetime import datetime, timedelta
```

● 55 ▲ 13 ▲

참여자—4

김종찬

서기현

유승균

이기중

## Naver 뉴스 기사 수집 자동화 (2020.12.05 ~ )

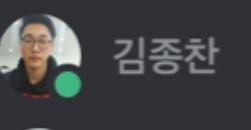
```
def get_urls(category='105'):
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.198 Safari/537.36'
    }
    date = (datetime.today() - timedelta(1)).strftime('%Y%m%d')
    last_p, urls = 1, []
    for page in range(1, 1000, 10):
        # 마지막 페이지로
        url = 'https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&listType=title&sid1={}&date={}&page={}'.format(category, date, page)
        req = requests.get(url, headers=headers)
        resp = TextResponse(req.url, body=req.text, encoding='utf-8')

        try:
            chk_next = resp.xpath('//div[@class="paging"]/a[@class="next nclicks(fls.page)"]/text()')[0].extract()
        except:
            chk_next = '끝'

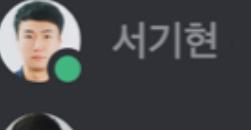
        if chk_next == '끝':
            # 마지막 페이지가 여러개일때
            # 마지막 페이지가 1개일때
            pages = resp.xpath('//a[@class="nclicks(fls.page)"]/text()' | 
                               '//div[@class="paging"]/strong').extract()
            last_p = pages[-1]
            print(last_p)
            break

        for page in range(1, int(last_p)+1):
            urls.append('https://news.naver.com/main/list.nhn?mode=LSD&mid=sec&listType=title&sid1={}&date={}&page={}'.format(category, date, page))
    return urls
```

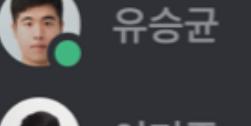
참여자—4



김종찬



서기현



유승균



이기중

Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

**출발하기**

Index +

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 자동화 (2020.12.05 ~ )

```
# Spider.py
%% writefile naver_news / naver_news / spiders / spider.py
import time
import re
import requests
import scrapy

from ..naver_articles import *
from scrapy.http import TextResponse
from naver_news.items import NaverNewsItem

class NaverNewsSpider(scrapy.Spider):
    name = 'naver_news'
    allow_domain = ["https://news.naver.com"]
    categ = { # '정치': '100',
        '101': '경제',
        '102': '사회',
        '103': '생활/문화',
        # '세계': '104',
        '105': 'IT/과학'}
    user_agent = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.198 Safari/537.36'

    def start_requests(self):
        all_category_urls = []
        for category in self.categ.keys():
            all_category_urls.append(get_urls(category))

        for urls in all_category_urls:
            for url in urls:
                yield scrapy.Request(url, callback=self.parse)
        time.sleep(10)
```

참여자—4

김종찬

서기현

유승균

이기중

## Naver 뉴스 기사 수집 자동화 (2020.12.05 ~ )

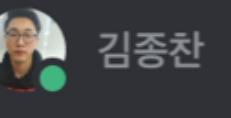
```
def parse(self, resp):
    links = resp.xpath('//*[@id="main_content"]/div[2]/ul/li/a/@href').extract()
    # links = [resp.urljoin(link) for link in links]
    for link in links:
        yield scrapy.Request(link, callback=self.parse_content)

def parse_content(self, resp):
    item = NaverNewsItem()
    title = resp.xpath('//*[@id="articleTitle"]/text() | //*[@id="content"]/div[1]/div/h2/text()')[0].extract()
    date = resp.xpath('//*[@id="main_content"]/div[1]/div[3]/div/span[@class="t11"]/text() | \
                       //div[@class="article_info"]/span[@class="author"]/em')[0].extract()
    content = resp.xpath('//*[@id="articleBodyContents"]/text() | \
                          //*[@id="articleBodyContents"]/strong/text() | \
                          //*[@id="articleBodyContents"]/div/text() | \
                          //*[@id="articleBodyContents"]/div/div/text() | \
                          //*[@id="articleBodyContents"]/font/text() | \
                          //*[@id="articleBodyContents"]/div[2]/ul/li/span/span/text() | \
                          //*[@id="articeBody"]/text()').extract()
    content = [text.replace('\xa0', ' ').strip() for text in content]
    try:
        c_num = resp.url.split('sid1=')[1].split('&')[0]

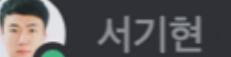
        item['date'] = re.findall('[0-9]{4}[.][0-9]{2}[.][0-9]{2}', date)[0]
        item['category'] = self.categ[c_num]
        item['press_agency'] = \
            resp.xpath('//a[@class="nclicks(atp_press)"]/img/@title | //div[@class="press_logo"]/a/img/@alt')[0].extract()
        item['link'] = resp.url
        item['title'] = title.strip()
        item['content'] = '\n'.join(content).strip()

        yield item
    except:
        print('nope') # url에 카테고리가 없는 연예기사 스포츠기사는 제외
```

참여자—4



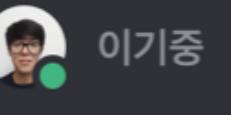
김종찬



서기현



유승균



이기중

### Index

#### # 프로젝트 동기 및 목표

#### # 진행 과정

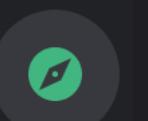
— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

#### # 결과



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 자동화 (2020.12.05 ~ )

```
# Pipelines.py
%%writefile naver_news/naver_news/pipelines.py
# %load naver_news/naver_news/pipelines.py

from itemadapter import ItemAdapter
from ..mongodb import collection
import time

class NaverNewsPipeline:
    def process_item(self, item, spider):
        # time.sleep(10)
        data = {
            'p_date': item['date'],
            'category': item['category'],
            'press_agency': item['press_agency'],
            'link': item['link'],
            'title': item['title'],
            'content': item['content'],
        }
        print('='*5)
        collection.insert(data)
        return item
```

참여자—4

김종찬

서기현

유승균

이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 수집 자동화 (2020.12.05 ~ )

```
# Setting.py
%%writefile naver_news/naver_news/settings.py
# %Load naver_news/naver_news/settings.py

BOT_NAME = 'naver_news'

SPIDER_MODULES = ['naver_news.spiders']
NEWSPIDER_MODULE = 'naver_news.spiders'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
#USER_AGENT = 'naver_news (+http://www.yourdomain.com)'

# Obey robots.txt rules
ROBOTSTXT_OBEY = False

# Configure maximum concurrent requests performed by Scrapy (default: 16)
#CONCURRENT_REQUESTS = 32

ITEM_PIPELINES = {
    'naver_news.pipelines.NaverNewsPipeline': 300,
}
```

참여자—4

김종찬

서기현

유승균

이기중

## Naver 뉴스 기사 중복제거

```
import pymongo
import pandas as pd
client = pymongo.MongoClient('mongodb://3.35.46.109:27017/')
db = client.news
collection = db.articles
items = collection.find()
df = pd.DataFrame(items)

from datetime import date, timedelta

yesterday = date.today() - timedelta(1)
date = yesterday.strftime('%Y.%m.%d')
df = df[df['p_date'] == date]
df['removal'] = date
df = df.drop(df.loc[df['content'] == ''].index)
df = df.drop(['_id'], axis=1)
df.dropna(inplace=True)
df.isnull().sum()
df.reset_index(drop=True, inplace=True)
```

참여자—4

김종찬

서기현

유승균

이기중

### Index



# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## Naver 뉴스 기사 중복제거

```
from sklearn.feature_extraction.text import TfidfVectorizer ..... TF-IDF 벡터 값을 구하는 라이브러리
from sklearn.metrics.pairwise import linear_kernel

tfidf = TfidfVectorizer()
tfidf_matrix = tfidf.fit_transform(df['content'])
n = len(df['content'])
cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)

num = []
i = 0
j = 0
l = []
cosine_sim[i][j]
for i in range(n - 1):
    for j in range(1, n - 1):
        if cosine_sim[i][j] >= 0.8:
            if i < j:
                num.append(j)
                l.append([i, j])
                #print(i, j)
                #print(cosine_sim[i][j])
new_ = []
for v in num:
    if v not in new_:
        new_.append(v)

article = df.drop(index=new_)

my_articles = article.to_dict('records')
client = pymongo.MongoClient('mongodb://3.35.46.109:27017/')
articles = client.news.articles
articles.insert(my_articles)
```

참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중

### Index

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## 데이터 저장 프로세스

### < 2년치 뉴스기사 scrapy >

| 각 날짜별 page URL scrapy  |  |   |
|--|--|---|
| Items.py   | spider.py  | 결과  |
| <p>&lt;수집 데이터 내용&gt;</p> <ul style="list-style-type: none"> <li>category = 카테고리</li> <li>date = 기사 날짜</li> <li>last_p = 마지막 페이지</li> </ul> | <ul style="list-style-type: none"> <li>start_requests = 각 날짜별 임의 page로 url 생성</li> <li>parse = 해당 날짜 마지막 페이지 체크</li> </ul> | <ul style="list-style-type: none"> <li>urls.csv</li> <li>각 카테고리별 CSV 파일 보유</li> </ul> |

### < Day scrapy >

| Items.py   | naver_articles.py   |
|--|---|
| <p>&lt;수집 데이터 내용&gt;</p> <ul style="list-style-type: none"> <li>category = 카테고리</li> <li>p_date = 기사 날짜</li> <li>press_agency = 언론사</li> <li>link = 기사 링크</li> <li>title = 기사 제목</li> <li>content = 기사 내용</li> </ul> | <ul style="list-style-type: none"> <li>전날 날짜의 끝페이지 확인</li> <li>카테고리 별, page 별 urls 생성</li> <li>return urls</li> </ul> |

### 기사 scrapy

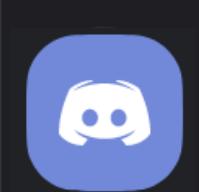
| Items.py  | spider.py   | 결과   |
|---|---|--|
| <p>&lt;수집 데이터 내용&gt;</p> <ul style="list-style-type: none"> <li>category = 카테고리</li> <li>p_date = 기사 날짜</li> <li>press_agency = 마지막 페이지</li> <li>link = 기사 링크</li> <li>content = 기사 내용</li> </ul> | <ul style="list-style-type: none"> <li>start_requests = urls.csv의 각 날짜와 마지막 페이지 생성</li> <li>parse = page의 기사리스트 link 크롤링</li> <li>parse_content = 기사 크롤링</li> </ul> | <ul style="list-style-type: none"> <li>articles.csv</li> <li>카테고리별 기사 CSV</li> </ul> |

### spider.py

| spider.py  |
|--|
| <ul style="list-style-type: none"> <li>start_requests = naver_articles.py를 통해 page 별 urls 가져옴</li> <li>parse = page의 기사리스트 link 크롤링</li> <li>parse_content = 기사 크롤링</li> </ul> |

### 결과

| 결과  |
|---|
| <ul style="list-style-type: none"> <li>pipelines.py</li> <li>MongoDB로 저장</li> </ul> |



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

- ~ Index +
- # 프로젝트 동기 및 목표
- # 진행 과정
  - 데이터 수집
  - 데이터 저장 및 관리
  - 챗봇 개발
  - 챗봇 테스트

## CSV에 저장

| _id      | p_date  | category | press_agency | link   | title  | content   |
|----------|---|----------|--------------|--|--|---|
| 0        | 2020.11.29  | IT/과학    | 이데일리         | https://news.naver.com/main/read.nhn?mode=LSD&...<br>[문자] 문마을 전 차관 과학기<br>설정적 연구원장 후보서 물러나야" | 공공연구노조 "문마을 전 차관 과학기<br>설정적 연구원장 후보서 물려나야"             | 지난 26일 내부 출신 인사 2명과 후보에 이를 끌어노는 성명서 통해 후보 사퇴...         |
| 1        | 2020.11.29  | IT/과학    | 블로터          | https://news.naver.com/main/read.nhn?mode=LSD&...<br>'신세계 정용진 정유경 남매 중여세<br>'3000억' 낸다       | 신세계 정용진 정유경 남매 중여세<br>'3000억' 낸다                       | 정용진 신세계그룹 부회장과 정유경 신세계 총<br>괄사장이 내야 할 중여세 규모가 3000...   |
| 2        | 2020.11.29  | IT/과학    | 서울경제         | https://news.naver.com/main/read.nhn?mode=LSD&...<br>LG, 인도 시장 출시…공략 키워드<br>는 '언택트'          | LG, 인도 시장 출시…공략 키워드<br>는 '언택트'                         | 30일 인도 시장 시장...비대면 마케팅 적극 활용<br>'n' 인도 주요 온라인 쇼핑몰 풀립... |
| 3        | 2020.11.29  | IT/과학    | 전자신문         | https://news.naver.com/main/read.nhn?mode=LSD&...<br>[부제] 국방성(부산일보 논설위원)씨 장<br>인상            | [부제] 국방성(부산일보 논설위원)씨 장<br>인상                           | ▲ 송도영씨 별세, 광명성(부산일보 논설위원)씨<br>장인상=28일, 대동 병원 장례식장 2...  |
| 4        | 2020.11.29  | IT/과학    | 블로터          | https://news.naver.com/main/read.nhn?mode=LSD&...<br>[블로터언박]교보문고 전자책 단말기<br>'Sam 7.8'        | [블로터언박]교보문고 전자책 단말기<br>'Sam 7.8'                       | 전자책(e-Book, 이북) 독서 인구가 해마다 늘고<br>있다고 한다. 지난 3년 문제...    |
| ...      | ...   | ...      | ...          | ...  | ...  | ...   |
| 438444   | 2019.01.01  | IT/과학    | 블로터          | https://news.naver.com/main/read.nhn?mode=LSD&...<br>[12월-4주] 주간 포털 브리핑                      | 이미 발표된 포털업계의 소식들을 모아 한눈에<br>볼 수 있도록 매주 보여드리고자 합니다...   |   |
| 438445   | 2019.01.01  | IT/과학    | 파이낸셜뉴스       | https://news.naver.com/main/read.nhn?mode=LSD&...<br>SK텔레콤, 사람과 로봇이 상동하는<br>5G 혁신 만들다        | SK텔레콤은 기해년(己亥年) 활급해지의 해를<br>맞아 미래의 주역인 SK텔레콤 신입사원...   |   |
| 438446   | 2019.01.01  | IT/과학    | 뉴스1          | https://news.naver.com/main/read.nhn?mode=LSD&...<br>[인사]정보통신산업진흥원(NIPA)                     | (서울=뉴스1) 남도영 기자 = ◆정보통신산업진<br>흥원(NIPA)이<br>경영전략실장>입니다. |   |
|          |   |          |              | https://news.naver.com/main/read.nhn?mode=LSD&...<br>SK그룹 주요 관계사 CEO 'CFS' 출범                | SK그룹 주요 관계사 최고경영자(CEO)가 'CFS'                          |   |
| In [7]:  | start = time.time()<br>df_it = pd.read_csv('naver_it_201129_190101.csv')<br>print(time.time() - start, 'sec')<br>df_it        |          |              |  |  |   |
| Out[7]:  | 10.588726043701172 sec  |          |              |  |  |   |
| In [8]:  | start = time.time()<br>df_cul = pd.read_csv('naver_cul_201129_190101.csv')<br>print(time.time() - start, 'sec')<br>df_cul     |          |              |  |  |   |
| Out[8]:  | 17.090404272079468 sec  |          |              |  |  |   |
| In [9]:  | start = time.time()<br>df_eco = pd.read_csv('naver_economy_201129_190101.csv')<br>print(time.time() - start, 'sec')<br>df_eco |          |              |  |  |   |
| Out[9]:  | 57.261399030685425 sec  |          |              |  |  |   |
| In [10]: | start = time.time()<br>df_soc = pd.read_csv("naver_society_201129_190101.csv")<br>print(time.time() - start, 'sec')<br>df_soc |          |              |  |  |   |
| Out[10]: | 85.34456443786621 sec   |          |              |  |  |   |

참여자-4

- 김종찬  
서기현  
유승균  
이기중



# MongoDB에 저장

Robo 3T - 1.4

File View Options Window Help

New Connection (4)

- > System
- > READ\_ME\_TO\_RECOVER\_YOUR\_DA...
- > config
- > news
  - Collections (1)
    - articles
  - Functions
  - Users

Welcome × db.getCollection('articles')

New Connection 3,35,46,109:27017 news

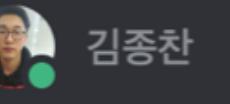
db.getCollection('articles').find({p\_date : '2020.12.09'})

articles 0.042 sec.

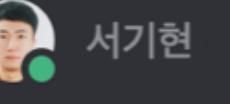
| _id | p_date         | category   | press_agency | link  | title                           | content                             |
|-----|----------------|------------|--------------|-------|---------------------------------|-------------------------------------|
| 1   | ObjectId("5... | 2020.12.09 | 경제           | 뉴스    | https://news.naver.com/main/... | 텐센트, 클라우드 사업 확대... 직원 2배 늘린다        |
| 2   | ObjectId("5... | 2020.12.09 | 경제           | 머니S   | https://news.naver.com/main/... | 이디야커피, 크리스마스 한정판 카드 출시              |
| 3   | ObjectId("5... | 2020.12.09 | 경제           | 헤럴드경제 | https://news.naver.com/main/... | 영 코로나19 백신 접종 하루만에 2명 '알레르기'        |
| 4   | ObjectId("5... | 2020.12.09 | 경제           | JTBC  | https://news.naver.com/main/... | 변창흠식 '반값 아파트' 검토... 다시 꺼낸 토지임대주택    |
| 5   | ObjectId("5... | 2020.12.09 | 경제           | 노컷뉴스  | https://news.naver.com/main/... | 다중대표소송제 감사위원 분리선출제 도입...상법 개정       |
| 6   | ObjectId("5... | 2020.12.09 | 경제           | 서울경제  | https://news.naver.com/main/... | '잘못 송금한 돈' 내년 7월부터 예보가 돌려준다         |
| 7   | ObjectId("5... | 2020.12.09 | 경제           | MBC   | https://news.naver.com/main/... | "많이 받은 사람 더 내라"...실손보험 대수술          |
| 8   | ObjectId("5... | 2020.12.09 | 경제           | 부산일보  | https://news.naver.com/main/... | 비대면 서비스 바우처 점검하니...재테크 취미교육 등 부적... |
| 9   | ObjectId("5... | 2020.12.09 | 경제           | 머니투데이 | https://news.naver.com/main/... | [부고] 문경훈씨(KB증권 센터장) 부진상             |
| 10  | ObjectId("5... | 2020.12.09 | 경제           | 머니S   | https://news.naver.com/main/... | 이마트24 PL 라면과 대포 숙취해소음로가 만났다! 속풀라... |
| 11  | ObjectId("5... | 2020.12.09 | 경제           | 뉴스    | https://news.naver.com/main/... | 한화그룹, 사회취약계층 1만 가구에 방역물품 전달         |
| 12  | ObjectId("5... | 2020.12.09 | 경제           | 뉴스    | https://news.naver.com/main/... | 경기 여주 메추리 농장서 고병원성 AI 확진            |
| 13  | ObjectId("5... | 2020.12.09 | 경제           | 이데일리  | https://news.naver.com/main/... | 휴젤, 임시주주총회 개최...동양에이씨씨 흡수합병 마무리     |
| 14  | ObjectId("5... | 2020.12.09 | 경제           | 부산일보  | https://news.naver.com/main/... | 고등과학원 오픈 교수, 한국인 최초 미국수학회 부회장 선출    |
| 15  | ObjectId("5... | 2020.12.09 | 경제           | 부산일보  | https://news.naver.com/main/... | [인사] 과학기술정보통신부                      |
| 16  | ObjectId("5... | 2020.12.09 | 경제           | MBN   | https://news.naver.com/main/... | '분노인증서' 내일부터 안 써도 된다..."배상책임은 강화해야" |
| 17  | ObjectId("5... | 2020.12.09 | 경제           | KBS   | https://news.naver.com/main/... | [집중취재]② 블록체인 금융중심지로...지원책 절실        |
| 18  | ObjectId("5... | 2020.12.09 | 경제           | 부산일보  | https://news.naver.com/main/... | 공정위, 가맹·대리점 포럼...대리점 보호제도 도입 논의     |
| 19  | ObjectId("5... | 2020.12.09 | 경제           | KBS   | https://news.naver.com/main/... | [집중취재]① 외국계 기업 첫 유치...금융중심지 10년' 결실 |
| 20  | ObjectId("5... | 2020.12.09 | 경제           | 부산일보  | https://news.naver.com/main/... | 석탄공사, '2020 공공기관 종합청렴도' 1등급 달성      |
| 21  | ObjectId("5... | 2020.12.09 | 경제           | 한국경제  | https://news.naver.com/main/... | '온라인' 채 입혔다...아모레퍼시픽 '화색'           |
| 22  | ObjectId("5... | 2020.12.09 | 경제           | 한국경제  | https://news.naver.com/main/... | 코카콜라·부킹...내년 유망주 10                 |
| 23  | ObjectId("5... | 2020.12.09 | 경제           | 연합뉴스  | https://news.naver.com/main/... | [그래픽] 코스피 지수 추이                     |
| 24  | ObjectId("5... | 2020.12.09 | 경제           | MBC   | https://news.naver.com/main/... | 내일 공인인증서 폐지...'공공·민간인증서' 모두 사용      |
| 25  | ObjectId("5... | 2020.12.09 | 경제           | 한국경제  | https://news.naver.com/main/... | 플랫폼 기업 집중 투자 ETF 뜬다는데...            |
| 26  | ObjectId("5... | 2020.12.09 | 경제           | 데일리안  | https://news.naver.com/main/... | 롯데마트, 가성비 품은 '롯데 시그니처' 와인 첫 선       |
| 27  | ObjectId("5... | 2020.12.09 | 경제           | 동아일보  | https://news.naver.com/main/... | 현대모비스, 출해 최고 지식재산경영기업 선정... 산업부장... |
| 28  | ObjectId("5... | 2020.12.09 | 경제           | 한국경제  | https://news.naver.com/main/... | 키움PE 대표에 김동준 선임                     |

Logs

참여자—4



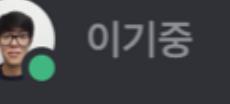
김종찬



서기현



유승균



이기중

## 챗봇 기본 템플릿

```
import discord ..... discord를 import 해주어야 실행이 됨
import pandas as pd
import nltk ..... 자연어처리를 위한 패키지
import pymongo

from wordcloud import WordCloud
from konlpy.tag import Mecab
from discord.ext import commands

token = "BOT_KEY" ..... https://discord.com/developers/applications에서 제공하는 Bot Token 입력

bot = commands.Bot(command_prefix='!')
bot.remove_command('help') ..... 기존 내장되어 있는 help 함수 제거

@bot.event
async def on_ready():
    print('Ready!!')

...
bot.run(token) ..... 코드 맨 마지막 부분에 있으며, 봇을 작동시킴
```

- 참여자—4
- 김종찬
  - 서기현
  - 유승균
  - 이기중

출발하기

Index

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## 챗봇 구성 함수

```
def database(p_date): ..... DB에서 추출하여 데이터프레임 만드는 함수
    client = pymongo.MongoClient('database')
    db = client.news
    ls = list(db.articles.find({'p_date': p_date}))
    df = pd.DataFrame(ls)
    df = df[df['removal'].notnull()].reset_index().drop(columns=['index']) ..... 중복기사 제거
    return df

def tdk(p_date): ..... 해당 날짜 키워드로 워드클라우드 만드는 함수
    mecab = Mecab() ..... 자연어처리
    df = database(p_date) ..... 데이터프레임 불러오기
    contents = [] ..... 기사 데이터 추출
    for data in df['content']:
        if type(data) == str:
            contents.append(data.strip())
    nouns = [] ..... 명사 단어 추출
    for idx in range(len(contents)):
        nouns.extend(mecab.nouns(contents[idx]))
    with open('ko_stopwords.txt', 'rt') as txt_file:
        stop_words = txt_file.readlines()
    stop_word = []
    for idx in range(len(stop_words)):
        stop_word.append(stop_words[idx].replace("\n", ""))
    new_nouns = [each_word for each_word in nouns if each_word not in stop_word] ..... 불용어 처리
    words = nltk.Text(new_nouns, name='words') ..... 단어 중복 체크
    data = words.vocab().most_common(100) ..... 상위 100개 단어 추출
    wordcloud = WordCloud(font_path='/usr/share/fonts/truetype/nanum/NanumMyeongjoExtraBold.ttf', ..... 워드클라우드 만들기
                           relative_scaling=0.05,
                           background_color='white',
                           ).generate_from_frequencies(dict(data))
    wordcloud.to_file('1.png')
    return data
```

참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중

뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

## Index

# 프로젝트 동기 및 목표

# 진행 과정

데이터 수집

데이터 저장 및 관리

챗봇 개발

챗봇 테스트

# 결과

## 챗봇 명령어 사용 ① - !summary 날짜

```
@bot.command()
async def summary(ctx, p_date):
    embed = discord.Embed( # 이미지 파일 불러오기 위한 엠베드 추가
        title='wordcloud',
        color=discord.Color.blue()
    )
    datas = tdk(p_date)
    number = 0 # 단어 총 갯수
    for data in datas:
        number += data[1]
    words = [] # % 구하기
    counts = [] # count 구하기
    for data in datas[:10]:
        words.append(data[0] + " " + str(round(data[1] / number * 100, 2)) + "%")
        counts.append(data[0] + " " + str(data[1]) + "개")
    wordstr = " ".join(words)
    countstr = " ".join(counts)
    file = discord.File("1.png", filename="image.png")
    embed.set_image(url="attachment://image.png")
    embed.add_field(name="word %", value=wordstr, inline=False) # %정보 엠베드에 붙이기
    embed.add_field(name="word_count", value=countstr, inline=False) # 정보 카운트
    await ctx.send(file=file, embed=embed) # 엠베드 보내기
```

해당 날짜의 뉴스 기사들 중 빈도수가 높은 단어를  
워드클라우드 이미지로 전송

### wordcloud

#### word %

기업 3.57% 코로나 2.5% 경제 2.44% 시장 2.39% 금지 2.25% 한국  
1.73% 투자 1.72% 금융 1.62% 서울 1.51% 정부 1.49%

#### word\_count

기업 7140개 코로나 5016개 경제 4890개 시장 4794개 금지 4509개 한  
국 3471개 투자 3435개 금융 3252개 서울 3021개 정부 2991개



## 사용 예시

!summary 2020.12.07

참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

## Index

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

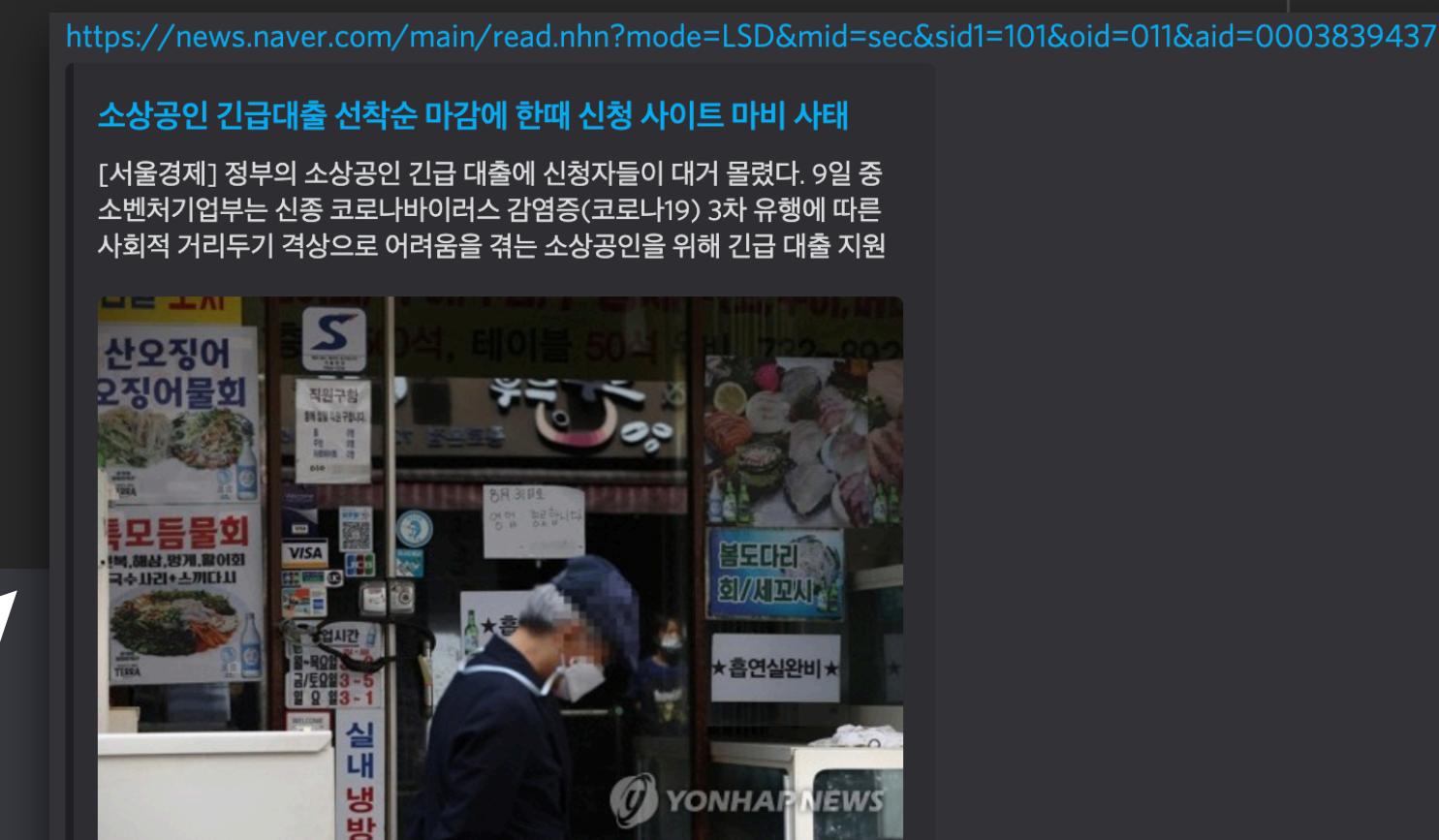
— 챗봇 테스트

# 결과

## 챗봇 명령어 사용 ② - !content 날짜 \*키워드

```
@bot.command()          날짜   키워드
async def content(ctx, p_date, *search):
    df = database(p_date)
    mecab = Mecab()
    articles = {}
    for data in df.iterrows():
        # 기사 하나씩 대조
        news = data[1].content
        # 단어
        words = nltk.Text(mecab.nouns(news))
        count = 0
        for idx in range(len(search)):
            count += words.vocab()[search[idx]]
        # 딕셔너리에 저장
        if count != 0:
            articles[data[1].link] = count
        # count 갯수로 정렬
    sorted_articles = dict(sorted(articles.items(), key=lambda item: item[1], reverse=True))
    limit = 0
    # 추천 기사
    rcd = []
    for link in sorted_articles.keys():
        rcd.append(link)
        limit += 1
        # 추천 갯수 제한
        if limit == 5:
            break
    try:
        for idx in range(len(rcd)):
            await ctx.send(rcd[idx])
    except:
        ctx.send('추천 기사를 모두 불러왔습니다!')
```

해당 날짜의 뉴스 기사들 중 기사 내용에 해당 키워드가  
가장 많이 들어 있는 순으로 뉴스 기사 링크 전송



### 사용 예시

!content 2020.12.07 인공지능 코로나

참여자—4

김종찬

서기현

유승균

이기중



뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

## Index

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## 챗봇 명령어 사용 ③ - !t\_pick 날짜

```
# 오늘의 키워드를 이용한 추천 기사
@bot.command()
async def t_pick(ctx, p_date):
    날짜
    df = database(p_date)
    datas = tdk(p_date)
    t_key = []
    for keyword in datas[:5]:
        t_key.append(keyword[0])
    mecab = Mecab()
    articles = {}
    for data in df.iterrows():
        news = data[1].content # 기사 하나씩 대조
        words = nltk.Text(mecab.nouns(news)) # 단어
        count = 0
        for idx in range(len(t_key)):
            count += words.vocab()[t_key[idx]]
        if count != 0: # 딕셔너리에 저장
            articles[data[1].link] = count
    sorted_articles = dict(sorted(articles.items(), key=lambda item: item[1], reverse=True)) # count 갯수로 정렬
    limit = 0
    rcd = [] # 추천 기사
    for link in sorted_articles.keys():
        rcd.append(link)
        limit += 1
        if limit == 5: # 추천 갯수 제한
            break
    try:
        for idx in range(len(rcd)):
            await ctx.send(rcd[idx])
    except:
        ctx.send('추천 기사를 모두 불러왔습니다!')
```

해당 날짜의 뉴스 기사 내용 중 빈도수가 가장 높은 단어들이 포함된 뉴스 기사 상위 5개 전송

<https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=022&aid=0003530570>

이자도 못내는 '줌비기업' 4000개... "구조조정 꼭 필요"

2019년 한계기업 사상 최대 "실업급여 기간 늘리고 정책 금융 줄여야" 1년 동안 영업해 대출 이자도 못 내는 한계기업이 지난해 4000개에 달한 것으로 집계됐다. 한계기업이 정상화·퇴출되지 않고 외부 지원으로 겨



## 사용 예시

!t\_pick 2020.12.07





## 챗봇 명령어 사용 ④ - !help

```
@bot.command()
async def help(ctx):
    embed = discord.Embed(
        title='명령 리스트!!(command list)',
        color=discord.Color.blue()
    )

    embed.set_thumbnail(
        url='https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQLByFMbwV_iorbq0iETHCYcuSjXbp7G4BMsA&usqp=CAU')
    embed.add_field(name="summary (date)", value='입력 날짜의 워드크라우드와 워드카운트를 보여줍니다.', inline=False)
    embed.add_field(name="content (date, *keywords)", value='입력한 키워드들을 바탕으로 입력된 키워드가 많이 들어간 기사를 찾아줍니다.', inline=False)
    embed.add_field(name="t_pick (date)", value='입력한 날짜에 가장 많이 쓰여진 키워드를 바탕으로 기사를 추천 해줍니다.', inline=False)

    await ctx.send(embed=embed)
```

### 명령어 리스트 정보 제공

#### 명령 리스트!!(command list)

##### summary (date)

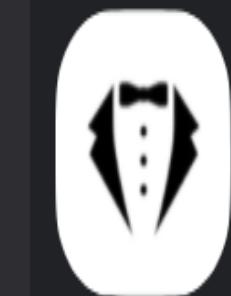
입력 날짜의 워드크라우드와 워드카운트를 보여줍니다.

##### content (date, \*keywords)

입력한 키워드들을 바탕으로 입력된 키워드가 많이 들어간 기사를 찾아줍니다.

##### t\_pick (date)

입력한 날짜에 가장 많이 쓰여진 키워드를 바탕으로 기사를 추천 해줍니다.



### 사용 예시

!help

## 느낀점

- 다양한 변수들로 인해 계획했던 것보다 시간이 오래걸려 시간 조절이 어려웠음
- 계획을 세울 때 좀 더 세분화하여 작업 소요시간 파악과 체계적인 계획수립으로 효율적인 시간활용이 되야할 것 같음
- 각 사이트마다 다양한 방식으로 크롤링을 하면서 크롤링 방식마다 크고 작은 속도차이를 경험해볼 수 있었음
- 챗봇을 만들게 되면서 자연어처리 같이 아직 배우지 않았던 것을 새롭게 공부해가며 목표를 달성하는 재미가 있었음

참여자—4

- 김종찬
- 서기현
- 유승균
- 이기중

Index

+

# 프로젝트 동기 및 목표

# 진행 과정

— 데이터 수집

— 데이터 저장 및 관리

— 챗봇 개발

— 챗봇 테스트

# 결과

## 추후 연구 및 개선 방향

- 추후 코사인 유사도를 좀 더 공부하여 중복제거에 힘을 쓸고 싶음
- 추천 알고리즘에 대해 호기심이 생겨 공부를 하고자 함
- 현재 **for**문으로는 ( $N=$ 기사수)  $N*N$ 번 중복체크로 메모리 부담이 많이 되는데, 메모리 부담이 적게 되도록 연구가 필요함
- 자연어처리에서 불용성 단어 리스트는 주기적으로 업데이트가 필요하며, 불용어 단어 처리 기준을 명확히 할 필요가 있음

## 느낀점

- 다양한 변수들로 인해 계획했던 것보다 시간이 오래걸려 시간 조절이 어려웠음
- 계획을 세울 때 좀 더 세분화하여 작업 소요시간 파악과 체계적인 계획수립으로 효율적인 시간활용이 되야할 것 같음
- 각 사이트마다 다양한 방식으로 크롤링을 하면서 크롤링 방식마다 크고 작은 속도차이를 경험해볼 수 있었음
- 챗봇을 만들게 되면서 자연어처리 같이 아직 배우지 않았던 것을 새롭게 공부해가며 목표를 달성하는 재미가 있었음

참여자—4

김종찬

서기현

유승균

이기중

## Index



## # 프로젝트 동기 및 목표

## # 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

## # 결과

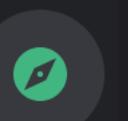
## 추후 연구 및 개선 방향

- 추후 코사인 유사도를 좀 더 공부하여 중복제거에 힘을 쓸고 싶음
- 추천 알고리즘에 대해 호기심이 생겨 공부를 하고자 함
- 현재 **for문**으로는 ( $N=기사수$ )  $N*N$ 번 중복체크로 메모리 부담이 많이 되는데, 메모리 부담이 적게 되도록 연구가 필요함
- 자연어처리에서 불용성 단어 리스트는 주기적으로 업데이트가 필요하며, 불용어 단어 처리 기준을 명확히 할 필요가 있음



#NewsBot에 메시지 보내기





뉴스 크롤링 프로젝트!  
멋지게 출발해 볼까요?

출발하기

Index +

# 프로젝트 동기 및 목표

# 진행 과정

- 데이터 수집
- 데이터 저장 및 관리
- 챗봇 개발
- 챗봇 테스트

# 결과

# Thanks News Bot !

[https://github.com/GIGI123422/crawl\\_prj](https://github.com/GIGI123422/crawl_prj)

참여자 - 4

김종찬

서기현

유승균

이기중